Issues in contemporary statistics and null hypothesis significance testing

Issues in Contemporary Statistics

Research is a fundamental part of the field of psychology, and has contributed immensely to its advances. Despite its contributions, however, research also comes with its own set of problems. These problems are evident in various steps of the research process, including study planning, research design, data collection, analysis, and reporting. During the planning stage, researchers are typically expected to determine the power of the study they are about to conduct, or the likelihood of detecting a statistically significant effect given that such effect truly exists in the population (Kline, 2008). Some of the factors that must be considered when estimating power are the hypothesized population effect size, the level of type I error and the size of the study sample (Cohen, 1990). Unfortunately, however, a-priori power analyses are not commonplace in psychological research. Although recommendations to report statistical power first appeared in the fourth edition of Publication Manual of the American Psychological Association, there are no examples on how it should be reported. This is in stark contrast to the different examples given by the fourth and fifth editions of the manual delineating how to report p-values (Gigerenzer, 2004). On a related note, according to Cohen (1990), the estimated power to detect a medium effect size in the literature is around 50%, or just about chance level. This is problematic because such level of statistical power means that, on average, results from these studies are just as good as guessing whether there is a significant effect. Although Cohen initially reported this level of statistical power in the 1962, it is not until very recently that researchers are beginning

to implement power analyses more frequently in their research (Gigerenzer, 2004).

Another problem in contemporary statistics involves the abundance of non-exploratory studies that include too many independent and dependent variables (Cohen, 1990). Just the sheer number of variables included in a study means that one will run many tests to examine the relation between these variables. This practice greatly increases type I error, making it more likely to detect a statistically significant effect when one does not truly exist. Such results may not be theoretically relevant or meaningful, but simply spurious. Although researchers have attempted to address the inflated type I error associated with conducting multiple tests, these methods are not ideal. For example, using post-hoc adjustment greatly reduces your power to detect a result if one truly exists, as you set your per-test significance criterion much lower than . 05. Unless the effect is large enough, it is unlikely to be detected with a very strict significance criterion. Because of that, the best practice is to investigate specific variables and research hypotheses that are relevant to the main idea of the researcher. However, it is common practice for researchers to conduct studies with a great number of independent and dependent variables, run many statistical tests, inflating their type I error, and then only report the statistically significant results, while pretending the other variables were never included and the other analyses were never conducted.

Other problems include some of the ways in which data is analyzed. For example, a researcher may want to dichotomize a continuous variable to infer something about people who have high or low anxiety, or to simply

make their analyses easier. However, this variable, along with many other variables in psychology are continuous and do not necessarily have a cut-off point when measured with a non-diagnostic scale. In reality, separating participants in artificial groups by turning a continuous variable into a categorical one greatly reduces the variance of the variable in question, and decreases its correlation with other variables (Cohen, 1990).

Issues Related to Null Hypothesis Significance Testing

Incorrect combination of Fisher's and Neyman-Pearson's theories

In addition to the problems discussed above, important issues in psychological research are related to null hypothesis significance testing (NHST). First, a major problem relates to a fundamental misunderstanding and amalgamation of Fisher's null hypothesis testing and the decision theory of Neyman and Pearson. Fisher's null hypothesis testing consists of testing a null hypothesis, which did not necessarily have to be a nil (i. e. effect of 0) hypothesis. Upon testing this null hypothesis, the researcher must report the exact level of significance, and not whether it is lower than a certain (i. e. p <. 05) significance level. According to Fisher, the significance level (p) was a property of the data, representing the likelihood of obtaining this data, assuming that the null hypothesis were true. Fisher suggested that this method should only be used when the researcher did not know much about the area of study (Gigerenzer, 2004). In this framework, Fisher did not define an alternative hypothesis. There was also no mention of statistical power, effect sizes, or type I and II errors. In contrast, Neyman-Pearson's decision theory consists of defining two hypotheses, defining one's type I and II error

rates, as well as sample size prior to running the experiment. If the data is inconsistent with the null hypothesis, the alternate hypothesis is accepted. These two approaches have been wrongfully combined into a NHST approach, which consists of setting up a null hypothesis, using 5% as the conventional cut-off to reject the null hypothesis and accept the research hypothesis. This process is repeated without considering whether p <0. 05 is appropriate for the specific research area (Gigerenzer, 2004). Despite the combination of Fisher's and Neyman-Pearson's different approaches, these two are not compatible. Fisher's approach focuses on the likelihood of obtaining your data (the p value), assuming that the null hypothesis were true. He emphasized the careful consideration of p-values according to the research context, and he did not suggest that these values should be used to accept or reject hypotheses. In contrast, Neyman-Pearson's approach is more relevant to situations where decision-making is involved (e. g. quality control), and the probability of falsely accepting or rejecting the null hypothesis is defined prior to the experiment. Fisher's significance level, or *p* is defined as a property of the data (i. e. likelihood of the data if the null hypothesis were true), while Neyman-Pearson's alpha is defined as a property of the test (rate of falsely rejecting the null hypothesis). Thus, alpha and p-values are not interchangeable (Lambdin, 2011).

Misunderstanding of the meaning of p-values

Because these two approaches have been incorrectly combined under NHST, there are widespread misconceptions about the meaning of p-values. Essentially, many students and researchers believe that p-values are more informative than they actually are (Gigerenzer, 2004). For example, some

common misconceptions include that p-values represent the probability that a type I error was committed. However, in a given experiment type I error either was or was not committed (Kline, 2008). The probability of type I error is related to long-term otucomes, if the same experiment was repeated many times. P-values have also been misinterpreted as representing the probability that the null hypothesis is true. However, this is incorrect, because the probability of the data if the null hypothesis were true is not interchangeable with the probability of the null hypothesis being true, given the data (Cohen, 1994). Another common misinterpretation of p-values is that they represent the probability of successful replication of the findings in future research. However, given the low statistical power of most studies, it is highly likely that the findings will not be replicated. Moreover, another common misconception consists of interpreting the rejection of the null hypothesis as support for the theory being tested. Unfortunately, however, rejecting the null hypothesis is not enough support for a theory. The researcher must also eliminate (i. e. control for) other possible explanations of the findings. Finally, psychology researchers always test the nil hypothesis, which states that the effect (e. g. mean difference, correlation) is always 0. However, no mean difference or correlation is always equal to exactly zero. Finding a statistically significant effect and rejecting the null hypothesis is always possible given a large enough sample because p-values are influenced by sample size (Cohen, 1994).

Problems stemming from NHST

Some problems stemming from the use of NHST include the encouragement of all-or-nothing thinking, where any results with $p <$. 05 are viewed as

significant, while anything with $p >$. 05 is viewed as non-significant (Field, 2018). Someone may interpret two effects that are practically the same as completely different (e. g. when p= 0. 049 compared to p= 0. 052). Even though this value is just a rule of thumb, and Fisher never recommended its use in every context, the convention of p <. 05 is pervasive within psychological research. A related issue is the confusion of statistical and practical significance. Statistical significance is not enough to infer practical significance, as the former can simply be the by-product of sampling error, or of a sample size that is too big. To infer practical significance, the magnitude of the effect should be considered in the context of other related findings in the research area. However, researchers often imply practical significance based on statistical significance only (Field, 2018).

Other issues that have been associated with NHST is neglecting to report effect sizes and confidence intervals. For example, fewer than half of the studies published in the Journal of Experimental Psychology between 2009 and 2010 reported the effect size associated with their analyses, and none reported a confidence interval for an effect size (Fritz, Morris, & Richler, 2012). It appears that there is still great room for improvement as to how frequently it is reported. Although it has been suggested that confidence intervals have the same problems as NHST (Abelson, 1997), because researchers check whether a zero is included in the interval, these nevertheless provide a range of how accurate an estimate is. However, they can be misused for the interest of the researcher by using different confidence levels in order to exclude zero from the interval.

Considering NHST in a broader context

Although much can be said about the drawbacks of NHST and other issues in contemporary statistics, it must be pointed out that many of these problems can be conceptualized as part of a bigger problem in science. Specifically, questionable practices in statistics (e. g. including too many variables in a study or interpreting statistically significant findings as practically significant) can be related to a publication bias in favour of statistically significant and “ original” findings. This is especially relevant to the field of psychology, as around 90% of the publications are of significant findings (Field, 2018). To achieve academic success and secure funding, a researcher needs to publish. Since significant findings are the mostly likely ones to get published, researchers are highly incentivized to engage in practices that make such findings more likely, and to selectively report only significant p-values. A related problem is the low rate of replication. Considering that original findings are more likely to get published (Lambdin, 2012), researchers do not have a high incentive to conduct replication studies. Specifically, approximately 1. 6% of psychology studies published between 1900 and 2012 are replication studies (Makel, Plucker, & Hegarty, 2012). This is a problem, because an integral part of the scientific process is the reproducibility of research to distinguish true effects from false positive findings.

Considering that NHST has been misused by researchers because of the incentive structure set up in the social sciences, NHST is not inherently wrong. Researchers should focus on using NHST in a more constructive way. Specifically, they should routinely include power estimates prior to conducting a study, as well as report effect size and confidence intervals in

their results (Cohen, 1994). When conducted this way, NHST should be viewed as simply one of the options available to researchers to explore their research questions. For example, other significance tests include tests of goodness of fit, which are carried out on the residual values in a dataset with the purpose of assessing how well a statistical model fits the data (Abelson, 1997). Different methods have their strengths and weaknesses depending on the context. Careful consideration and critical thinking should be used in statistics, instead of the mechanical employment of NHST only.

References

- Abelson, R. (1997). On the Surprising Longevity of Flogged Horses – Why There Is a Case for the Significance. *Psychological Science* , 8, 12-15. doi: 10. 1111/j. 14679280. 1997. tb00536. x
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist* , *45* , 1304-1312. doi: 10. 1037/0003-066x. 45. 12. 1304
- Cohen, J. (1994). The earth is round, $p < .05$. *American Psychologist* , *49* , 997-1003. doi: 10. 1037/0003-066x. 49. 12. 997
- Field, A. (2018). *Discovering statistics using IBM SPSS* Statistics *(5 th Ed.)* . Thousand Oaks, CA: Sage
- Fritz, C. O., Morris, P. E., & Richler, J. J. (2012). Effect size estimates: Current use, calculations, and interpretation. *Journal of Experimental Psychology: General* , 141, 2–18. doi: 10. 1037/a0024338
- Gigerenzer, G. (2004). Mindless Statistics. *Journal of Socio-Economics* , 33, 587-606. doi: 10. 1016/j. socec. 2004. 09. 033
- Kline, R, B. (2009). *Becoming a behavioral science researcher: A guide to producing research that matters* . New York: Guilford Press

- Lambdin, C. (2012). Significance tests as sorcery: Science is empirical —significance tests are not. *Theory & Psychology* , 22 , 67–90. doi: 10. 1177/0959354311429854
- Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in Psychology Research. *Perspectives on Psychological Science* , *7* , 537–542. doi: 10. 1177/1745691612460688